Pangea: a platform for viral discovery spanning wastewater, urban centers, and astronauts

Christopher E. Mason[1]
Jonathan Foox[1]
Maria Sierra[1]
JJ Hastings[1]
Eliah Overbey[1]
Krista Ryon[1]
Braden Tierney[1]
Mark Sharkey[2]
Alejandro Mantero[2]
Naresh Kumar[2]
George Grills[2]
The MetaSUB Consortium[1],
Kasthuri J Venkateswaran[3]
Natasha Schaefer Solle[2]
Helena Solo-Gabriele[2]

1Department of Physiology and Biophysics Weill Cornell Medicine, New York, NY, 10065
[2]Department of Environmental Engineering, University of Miami, FL, 33136
[3]Biotechnology and Planetary Protection Group, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA,

Widespread deployment of sequencing technologies has led to improved discovery and quantification of known and emerging viruses across many environments, including municipal systems like wastewater and subways, animal reservoirs such as bats and peridomestic pets, and even in unique, selective environments like the International Space Station (ISS). Here, we describe a large-scale resource (Pangea) for RNA and DNA shotgun sequencing, annotated sampling, and computational tool set (mapping, assembly, characterization) for viral and microbial discovery. Our data set spans 60 cities from seven continents, continual wastewater sampling in targted cities, and multiple spacecraft in Low Earth Orbit (LEO), including the ISS and high-orbit samples from missions such as the Inspiration4. We used over 3,000 metatranscriptomic and amplicon sets from urban centers and wastewater to track the emergence and dominance of various SARS-CoV-2 variants (e.g. Alpha, Delta) over 2020-2021 in Florida, New York, and Wisconsin, 580 metatranscriptomic samples from feral and peridomestic cats to find new animal reservoirs of viruses, and 1,100 metagenomic samples from spacecraft before, during, and after LEO spaceflight to examine the impact of spaceflight on microbial ecology.

These results revealed a complex set of interactions of host and microbial responses in all systems. Our data led to the discovery of over 12,000 new viruses and bacteria in urban centers, as well as 838,532 CRISPR arrays not found in reference databases. The samples and data interface also revealed the emergence and colonization of microbes in spacecraft, with a concomitant response of T-cell expressed motifs (TCEMs) in astronauts that showed increasing response and exchange of the microbial peptides during missions, especially between crew members. Finally, our integrated clinical and environmental data showed a high concordance (R2=0.7) of patient SARS-CoV-2 prevalence and copies per L (cp/L) in wastewater, with an ability to predict the emergence of the COVID-19 case spike in patients several days early through use of the wastewater. Finally, we built a k-mer map of all the microbes and sequences found across all known databases on Earth and from LEO (MetaGraph), which enables an easy means to explore mapping, alignment, and assembly data, as well as create a fully reproducible and annotated database for data sharing, mining, and viral quantification.